Expert Systems With Applications 63 (2016) 97–111

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Review

A survey on using domain and contextual knowledge for human activity recognition in video streams

Leonardo Onofri[a], Paolo Soda[a,*], Mykola Pechenizkiy[b], Giulio Iannello[a]

[a] Department of Engineering, University Campus Bio-Medico of Rome, Via Alvaro del Portillo 21, 00128 Roma, Italy
[b] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven 5600 MB, The Netherlands

ARTICLE INFO

Article history:
Received 16 July 2015
Revised 29 February 2016
Accepted 8 June 2016
Available online 16 June 2016

Keywords:
Human activity recognition
Computer vision
A priori knowledge
Contextual information
Reasoning

ABSTRACT

Human activity recognition has gained an increasing relevance in computer vision and it can be tackled with either non-hierarchical or hierarchical approaches. The former, also known as single-layered approaches, are those that represent and recognize human activities directly from the extracted descriptors, building a model that distinguishes among the activities contained in the training data. The latter represent and recognize human activities in terms of subevents, which are usually recognized my means of single-layered approaches. Alongside of non-hierarchical and hierarchical approaches, we observe that methods incorporating a priori knowledge and context information on the activity are getting growing interest within the community. In this work we refer to this emerging trend in computer vision as knowledge-based human activity recognition with the objective to cover the lack of a summary of these methodologies. More specifically, we survey methods and techniques used in the literature to represent and integrate knowledge and reasoning into the recognition process. We categorize them as statistical approaches, syntactic approaches and description-based approaches. In addition, we further discuss public and private datasets used in this field to promote their use and to enable the interest readers in finding useful resources. This review ends proposing main future research directions in this field.

1. Introduction

Human activity recognition in video streams is an active research area presenting some of the most promising applications of computer vision such as network-based surveillance, content-based video analysis, user-interface and elderly monitoring. Network-based surveillance systems provide interactive, real-time monitoring which increases human efficiency and accuracy, especially with the growing number of cameras (Lin, Sun, Poovandran, & Zhang, 2008; McKenna, 2003; Niu, Long, Han, & Wang, 2004). Content-based video analysis and automatic annotation permit efficient searching, e.g. finding tackles in soccer matches or typical dance moves in music videos (Chang, 2002; Dimitrova, 2003; Hanjalic, Lienhart, Ma, & Smith, 2008). In the user-interface application domain, activity recognition can complement speech recognition and natural language understanding for helping in creating computers that can better interact with humans (Choi, Cho, Han, & Yang, 2008; Pentland, 1998; Shang & Lee, 2011). Finally, monitoring systems which recognize activities of daily living (ADL) can be applied to home care technologies for elderly, reducing the costs and burdens of care-giving while increasing safety and autonomy in old age (Cardinaux, Bhowmik, Abhayaratne, & Hawley, 2011; Khan & Sohn, 2011; Zouba, Boulay, Bremond, & Thonnat, 2008).

The general task of human activity recognition consists in labelling videos that contain human motion with activity classes. To this aim, activity recognition systems cope with a variety of issues, which depend on factors such as the type of acquired videos, the number of persons involved in the activity, the complexity of performed activities and so on. Moreover, these systems could face related topics such as human detection, human movement tracking and person identification, that might be used as lower level modules of an activity recognition system.

The recognition of human activities can be performed at various levels of abstraction. Hence, the goal of an activity recognition system may comprise, for instance, simple movements like "left leg forward" or "arm stretching"; higher complex movements like "running" or "handshaking"; compositions of low-level movements like "jumping hurdles" or "table clearing". One of the earliest attempt to propose a general definition of human motion was performed by Bobick (1997). He defined a *movement* as the most atomic human motion, an *activity* as a sequence of movements and an *action* as a large-scale event, typically including

[*] Corresponding author. Fax: +39 06 225419609.
E-mail addresses: leonardonofri@gmail.com (L. Onofri), p.soda@unicampus.it (P. Soda), m.pechenizkiy@tue.nl (M. Pechenizkiy), g.iannello@unicampus.it (G. Iannello).

http://dx.doi.org/10.1016/j.eswa.2016.06.011
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

interaction with the environment. Conversely, Turaga, Chellappa, Subrahmanian, and Udrea (2008) defined an *action* as a simple motion pattern usually executed by a single person and typically lasting for short durations of time, whereas an *activity* is a complex sequence of actions performed by several humans who could interact with one another. Moreover, Poppe (2010) adopted a hierarchical scheme as well, defining three levels of abstraction: the lowest level is named the *action primitive*, an *action* is a composition of action primitives that describes a whole-body movement and the *activity* contains a number of actions with a high-level interpretation of the movement.

The aforementioned definitions contain some evident inconsistencies. To avoid any confusion in terminology, we use the term *activity recognition* as the general motion categorization framework, irrespective of the abstraction level actually investigated. When a classification system deals with simple activities that do not show any hierarchy, there is no reason to introduce different definitions and we just use the term activity. On the contrary, when focusing on high level motion understanding, where the approaches typically rely upon a certain degree of hierarchy, following (Aggarwal & Ryoo, 2011) we use the concepts of *event* and *subevents*. A subevent is the lower level movement that is to be recognized, wherein the final goal is the recognition of a higher level activity (the event). For example, we will use the term subevent for the "left leg forward" movement, where the goal is to recognize the event "running", whereas we will use the term subevents for the "running" and "jumping" movements, where the goal is to recognize the event "jumping hurdles". Note that while referring to this kind of high-level activity we will often use the term of composite activities for stressing their property of being characterized by an event composed of subevents.

In a video the information is conveyed in the form of spatiotemporal pixel intensity variations and thus, extracting a suitable set of descriptors is an important prerequisite of any activity recognition system. Once they have been extracted and a set of class labels has been defined, human activity recognition can be formulated as a classification problem that can be tackled with either non-hierarchical or hierarchical approaches (Aggarwal & Ryoo, 2011; Vishwakarma & Agrawal, 2013).

The former, also known as single-layered approaches, are those that represent and recognize human activities directly from the extracted descriptors, building a model which distinguishes among the activities contained in the training data. Single-layered approaches are most effective when a pattern describing an activity can be captured from training sequences; these approaches are suitable for the recognition of gestures and actions, such as relatively simple (and short) sequential movements of humans (e.g., walking, jumping, and waving) (Gorelick, Blank, Shechtman, Irani, & Basri, 2007; Poppe, 2010; Schuldt, Laptev, & Caputo, 2004).

The latter represent and recognize human activities in terms of subevents, which are usually recognized my means of single-layered approaches. Hierarchical methodologies are able to recognize high-level activities because of their ability to incorporate knowledge on the activity structure, making the recognition process conceptually understandable and computationally tractable.

Alongside of non-hierarchical and hierarchical approaches, we observe that methods incorporating *a priori knowledge* and *context information* on the activity (see Section 2 for their definition) are getting growing interest in the literature. In this work we refer to this emerging trend in computer vision as *knowledge-based human activity recognition* (KBAR) with the objective to cover the lack of a summary of these methodologies. More specifically, we survey methods and techniques used to represent and to integrate knowledge and reasoning into the recognition process, whereas we do not focus on low-level modules such as body structure analysis, tracking and feature extraction.

The paper is organized as follows: Section 2 discusses the exploitable knowledge, whereas Section 3 overviews the approaches for knowledge-based exploitation in human activity recognition. Section 4 present the available datasets for testing the methodologies. Section 5 discusses the surveyed contributions, whereas Section 6 provides future directions and concludes the paper.

1.1. Comparisons with previous reviews

Previous reviews on human activity recognition have focused on different aspects of motion understanding. Bobick (1997) described different approaches dividing his analysis in three different levels of abstraction, i.e. movements, activities and actions. Aggarwal and Cai (1999) and Wang, Hu, and Tan (2003) discussed body structure analysis, tracking and recognition. Kruger, Kragic, Ude, and Geib (2007) reviewed human action recognition approaches while classifying them on the basis of the complexity of features involved in the action recognition process. Their reviews focused especially on the planning aspect of human action recognitions, considering their potential application to robotics. Poppe (2010) considered image representation and video classification, limiting his survey to simple activity recognition. Turaga et al. (2008) and Aggarwal and Ryoo (2011) focused on both simple and complex human activities, describing different approaches in terms of feature extraction and classification algorithms. In their paper, approaches are categorized on the basis of the complexity of the activities and in terms of the recognition methodologies they use. Vishwakarma and Agrawal (2013) and Suriani, Hussain, and Zulkifley (2013) directed their surveys towards surveillance systems. The former offers a summary for activity recognition in video surveillance, integrating the surveyed papers presented in Aggarwal and Ryoo (2011), and providing a discussion on object tracking. The latter focused on frameworks used in sudden event recognition, defined as a subset of an abnormal event in video surveillance applications, reporting also the requirements and a comparative studies of a sudden event recognition system. Recently, Ziaeefard and Bergevin (2015) surveyed methodologies for activity recognition in still images and videos using semantic features. The review identifies the pose, the poselet, the objects, the scene, and the attributes as semantic features and it mostly discusses how they can be extracted and used to recognize the human activities. It mentions that hierarchical representation and reasoning mechanisms can be used to recognize the activities, and it briefly discusses potential applications where semantic approaches may be of assistance. Nevertheless, this work does not address how knowledge needed to exploit semantic information can be represented and integrated into the recognition process.

2. Exploitable knowledge

Knowledge exploitation is an established approach in the data mining literature, since it is helpful for selecting suitable classification techniques, pruning the space of hypothesis and improving the overall performance (Nigro, Císaro, & Xodo, 2008). The several advantages of knowledge exploitation can be summarized as follows (Crevier & Lepage, 1997):

- With an explicit knowledge arrangement, data contradictions and omissions become apparent, thus suggesting alternative means of extracting information from videos and images.
- Knowledge-based techniques permit to design and develop in an intuitive (visual) manner the recognition system and to extract information from examples.
- Explicit knowledge representation allows the separate description and the parallel use of knowledge pertaining to different domains, such as knowledge about image processing, knowl-

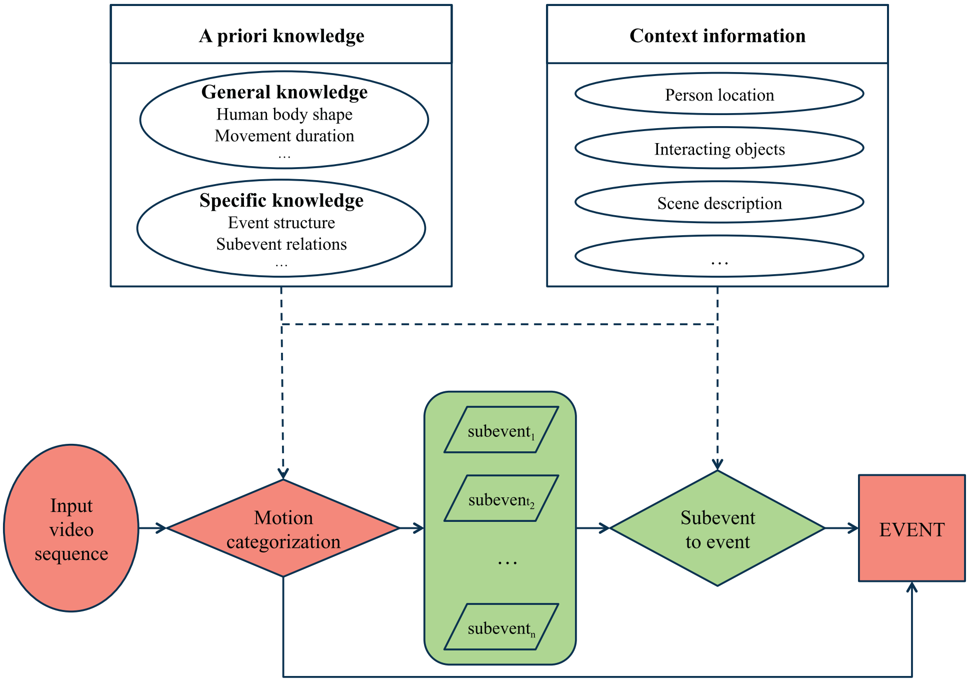

Fig. 1. Schematic of KBAR approaches including a priori knowledge and context information.

edge about the physical world, and knowledge about the specific task domain.

- Knowledge representation can make the computer vision algorithms more efficient, revealing the most suitable procedure that should be employed in a given situation.

To discuss how these advantages have been exploited in human activity recognition we find convenient to distinguish two types of knowledge: *a priori knowledge* and *context information*.

A priori knowledge includes all the general information deriving from the rationality of human activities, and it consists of both general and domain specific knowledge about the entities and the structure of any human activity. For instance, general knowledge of human body shape can help in extracting the silhouette of the person in the scene, whereas specific knowledge of the sequential structure of an event can help in recognizing its composing subevents.

Context information was defined by Dey as any relevant information that can be used to characterize the situation of an entity (Dey, 2001). With reference to human activities, examples of context information are: where you are, who you are with, what resources are nearby.

Fig. 1 shows that, when prior knowledge and contextual information are available, it is possible to incorporate these beliefs directly into the system to guide the recognition process. Red blocks in the figure represent the usual pipeline for simple activity recognition, where the system gathers the performed event directly from the input video sequence. On the contrary, KBAR approaches are typically endowed with green blocks since they determine an event on the basis of the recognized subevents. Within this process, knowledge exploitation can support both the recognition of the subevents and the categorization of the event given a sequence of subevents. Indeed, on the one hand, temporal and contextual relations among subevents, entities and environment support the recognition of a sequence of simple activities. On the other hand, domain specific knowledge can support the event modelling process, providing a way to infer the event, given the recognized lower level subevents. In order to clarify these assertions, we make use of two examples.

First, assume that an activity recognition system is aware that a subevent *A* may be followed by another subevent *B* with high probability, whereas the subevent *C* rarely occurs after *A*. This kind of a priori knowledge can be used to adjust the motion categorization step of Fig. 1 after that a subevent *A* has been recognized. The procedure can be easily generalized to a sequence of subevents allowing a more robust classification of composite activities.

As a second example, assume that a surveillance system is trained to recognize a set of security violations in a supermarket. Assume also that the system use a hierarchical strategy, so that it first recognizes subevents such as walking, bending, hand waving, etc. When the system analyses the recognized subevents to infer the performed event (subevent-to-event step of Fig. 1), the context information given by the opening time of the supermarket plays an important role in distinguishing among the events. Indeed, to detect a theft event in the supermarket when it is closed, it is sufficient to detect whether there is someone performing a subevent like picking up merchandise. Conversely, during the normal opening hours, the same subevent does not entail the theft event.

3. Knowledge-based approaches in human activity recognition

We categorize the knowledge representation and reasoning techniques used in the KBAR framework into three main categories: statistical approaches, syntactic approaches and description based approaches. It is worth noting that such a categorization was presented in Aggarwal and Ryoo (2011) to classify the hierarchical approaches for human action analysis in computer vision, and it was adopted also in Vishwakarma and Agrawal (2013). Although not all the papers surveyed here can be classified as hierarchical we maintain this categorization for two reasons. First, it allows an easy reference to recent surveys, which offer a different perspective of the field from ours. Second, even the non-hierarchical KBAR methods considered here can be naturally assigned to one of these categories.

Within each category we first give details on how it represents knowledge and/or enables reasoning. Further information on the specific approach is then given. Finally, we present peculiar-

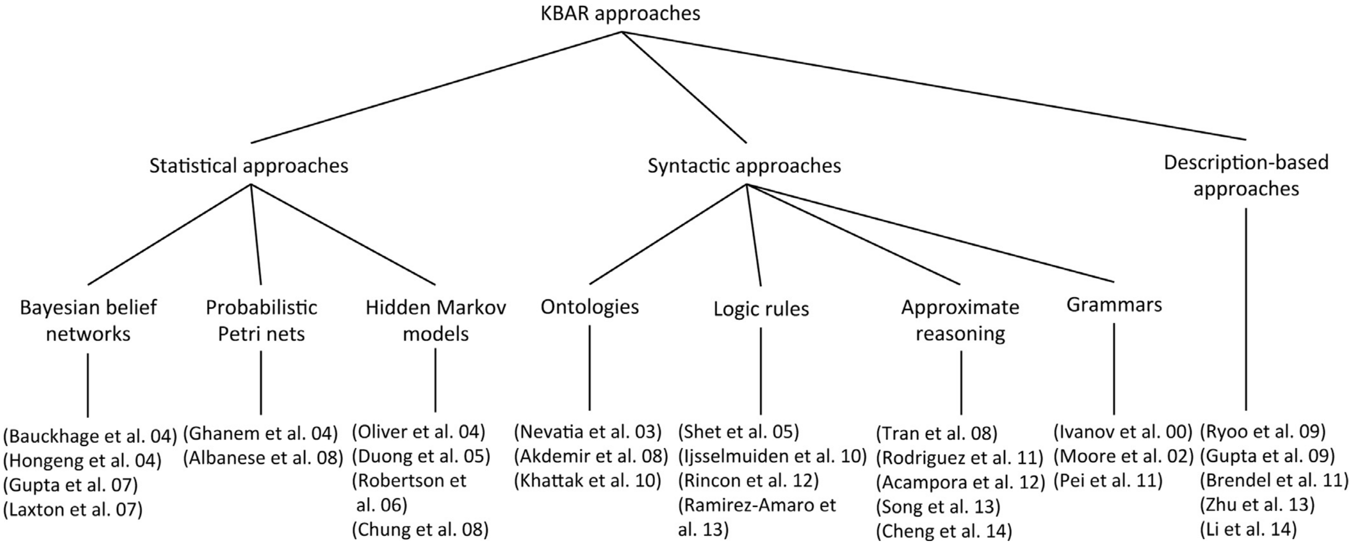

Fig. 2. Detailed taxonomy for KBAR approaches and the lists of publications corresponding to each category.

Table 1
Approaches that recognize human activities by means of prior knowledge, context information and reasoning techniques.

Paper	Data fusion	Hierarchical description	Context information	Knowledge and reasoning engine
(Ivanov & Bobick, 2000)	×	✓	Object detection	CFG
(Nevatia et al., 2003)	×	✓	Object detection and location	Ontologies
(Moore & Essa, 2002)	×	✓	Object detection	CFG
(Bauckhage et al., 2004)	✓	×	Object detection	BBN
(Ghanem et al., 2004)	×	✓	Object detection	PPN
(Hongeng et al., 2004)	×	✓	Spatial map	BBN
(Oliver et al., 2004)	✓	✓	Audio, keyboard and mouse activities	HMM
(Duong et al., 2005)	×	✓	×	HMM
(Shet et al., 2005)	×	✓	×	Logic rules
(Robertson & Reid, 2006)	×	✓	Location	HMM
(Gupta & Davis, 2007)	×	×	Object detection	BBN
(Laxton et al., 2007)	×	✓	Object detection	BBN
(Akdemir et al., 2008)	×	×	Spatial map	Ontologies
(Albanese et al., 2008)	×	✓	Object detection	PPN
(Chung & Liu, 2008)	×	✓	Spatial map	HMM
(Tran & Davis, 2008)	×	✓	Object detection and location	Approximate reasoning
(Gupta et al., 2009)	×	✓	×	Description-based
(Ryoo & Aggarwal, 2009)	×	✓	×	Description-based
(Ijsselmuiden & Stiefelhagen, 2010)	✓	×	Location, identity, visual focus of attention, speech and head pose	Logic rules
(Khattak et al., 2010)	✓	✓	Location, sound and time	Ontologies
(Brendel & Todorovic, 2011)	✓	✓	Object detection	Description-based
(Pei et al., 2011)	×	✓	Object detection	CSG
(Rodriguez-Benitez et al., 2011)	×	✓	Object detection	Approximate reasoning
(Acampora et al., 2012)	×	✓	Object detection and location	Approximate reasoning
(Rincón et al., 2013)	×	✓	×	Logic rules
(Song et al., 2013)	✓	✓	Object detection	Approximate reasoning
(Zhu et al., 2013)	×	✓	×	Description-based
(Ramirez-Amaro et al., 2013)	×	✓	Object detection	Logic rules
(Li & Fu, 2014)	×	✓	Object detection	Description-based
(Chen et al., 2014)	×	✓	×	Approximate reasoning

ities (e.g., integration of several input sources or multiple actor handling) of the single surveyed paper in chronological order. For the sake of completeness, Fig. 2 illustrates a taxonomy of the approaches, whereas Table 1 lists them. Column 2 of the table indicates if the method uses other sources than video sequences to extract information useful for activity recognition, column 3 shows whether the method uses a hierarchical description for composite events, column 4 describes the contextual information, if any, whereas column 5 shows the strategies used for knowledge representation and reasoning.

Since pointing out the datasets used in the experimental part is also of interest, Tables 2 and 3 list public and private datasets currently used by the surveyed papers (Section 4).

Table 2
Characteristics of public datasets used to test KBAR approaches. The datasets are listed in chronological order.

Dataset	Locus	Object	Actors	Camera	# Activities	Activity description	Other info	KBAR paper
Bank (Vu et al., 2003)	Indoor	✓	Multi	Single fixed, static background	2	Bank robberies and normal activities	Six video segments, four on bank robberies and two with normal activities. Length: 15-20 *s* per video.	(Akdemir et al., 2008; Albanese et al., 2008)
PETS 04 (Fisher, 2004)	Indoor, outdoor	✓	Multi	Static, wide angle camera lens, 384 × 288 pixels, 25 fps, MPEG2 compressed	6	Walking, browsing, collapse, leaving object, meeting, fighting	Public space surveillance task, 28 videos with ~ 26500 labeled frames	(Acampora et al., 2012; Chen et al., 2014)
TSA Airport Surveillance (Vaswani et al., 2005)	Outdoor	✓	Multi	Single fixed, 320 × 240 pixels	n.d.	Specific actions, e.g. plane take-off and landing, passenger transfer to/from the terminal, baggage loading and unloading, refueling	1 video, 118 min long	(Akdemir et al., 2008; Albanese et al., 2008)
IXMAS (Weinland et al., 2006)	Indoor	✓	Single	5 static cameras, 390 × 291 pixels, 23 fps, png compressed	11	Check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave hand, punch, kick, pick up	Each action is performed 3 times by 10 actors (5 males, 5 females)	(Rincón et al., 2013)
ViSOR (Vezzani & Cucchiara, 2007)	Indoor, outdoor	✓	Multi	Dynamic and fixed	n.d.	n.d.	Large set of multimedia data with physical objects, action/events, context information	(Rodriguez-Benitez et al., 2011)
Olympic Sports (Niebles et al., 2010)	Indoor, outdoor	✓	Single[a]	YouTube sequences, with severe occlusions, camera movements, compression artifacts	16	High jump, long jump, triple jump, pole vault, discus throw, hammer throw, javelin throw, shot put, basketball layup, bowling, tennis serve, platform (diving), springboard (diving), snatch (weightlifting), clean and jerk (weightlifting) and vault (gymnastics)	50 sequences per class	(Brendel & Todorovic, 2011)

(*continued on next page*)

Table 2 (*continued*)

Dataset	Locus	Object	Actors	Camera	# Activities	Activity description	Other info	KBAR paper
VIRAT Ground Video - Release 1 (Oh et al., 2011)	Outdoor	✓	Multi	Ground camera; 1920 × 1080 pixels	6	Loading, unloading, opening trunk, closing trunk, getting into vehicle, getting out of vehicle	25 hours 16 scenes	(Zhu et al., 2013)
VIRAT Ground Video - Release 2 (Oh et al., 2011)	Outdoor	✓	Multi	Aerial mobile camera; 640 × 480 pixels	11	Person loading an object to a vehicle, person unloading an object from a vehicle, person opening a vehicle trunk, person closing a vehicle trunk, person getting into a vehicle, person getting out of a vehicle, person gesturing, person carrying an object, person running, person entering a facility, person exiting a facility	4 hours, 1 scene	(Zhu et al., 2013)
WaRo11 (Santofimia et al., 2012)	Indoor	✓	Single	Fixed camera, single room	12	Check watch,cross arms,scratch head,sit down, get up, turn around, walk, wave hand, punch, kick, point, pick up	11 sequences, 1 per person, 5 min each	(Rincón et al., 2013)
MPII-Cooking (Rohrbach, Amin, Andriluka, & Schiele, 2012)	Outdoor	✓	Single	Single static camera, 1624 × 1224, 29.4fps	14	Make sandwich, salad, fried potatoes, potato pancake, omelet, soup, pizza, casserole, mashed potato, snack plate, cake, fruit salad, cold drink, and hot drink	44 videos with a total length of more than 8 hours (881755 frames)	(Li & Fu, 2014)

[a] The athlete is in the foreground, the crowd or other people in the background and do not take part in the sport activity.

Table 3
Private datasets used to assess the human activity recognition performances of KBAR approaches. The datasets are listed in chronological order. For space reason the acronym "D.-b." stands for description-based.

Approach	Paper	Locus	Actors	Camera	# Classes	Activity description
Statistics	(Bauckhage et al., 2004)	Indoor	Single	640 × 480	7	Drinking from a cup, reading a book, phoning, typing on the keyboard, etc.
	(Ghanem et al., 2004)	Outdoor	Multi	n.d.	8	Human car interaction in parking lot; primitive events: appears, disappears, moves, stops, enters-car, exits-car, enters-area, exits-area
	(Hongeng et al., 2004)	Outdoor	Multi	n.d.	n.d.	Examples: a person drops off a package, a person (a car) follows another person (another car)
	(Oliver et al., 2004)	Indoor	Multi	n.d.	6	Phone conversation, face to face conversation, presentation, other activity, nobody around, distant conversation
	(Duong et al., 2005)	Indoor	Single	5 cameras	6	Entering-the-room, making breakfast, eating breakfast, washing-dishes, making-coffee, reading morning newspaper, having coffee, leaving the room
	(Robertson & Reid, 2006)	Outdoor	Multi	n.d.	Surveillance: 8 Sport: 33	Surveillance: walking, running, stopping in five directions; Sport: 33 tennis strokes
	(Gupta & Davis, 2007)	Indoor	Single	n.d.	6	Drinking, spraying, answering a phone call, making a phone call, pouring from a cup, lighting the flashlight
	(Laxton et al., 2007)	Indoor	Single	RGB and depth camera	30	Cooking activities, e.g. crack egg, cook bread
	(Chung & Liu, 2008)	Indoor	Single	n.d.	5	Walk to toilet, take a walk, watch TV, go to eat meal, take a shower
Syntactic	(Ivanov & Bobick, 2000)	Outdoor	Multi	n.d.	n.d.	Parking lot monitoring task
	(Nevatia et al., 2003)	Outdoor	Multi	n.d.	1	Probability of the occurrence of event "stealing by blocking"
	(Moore & Essa, 2002)	Indoor	Multi	n.d.	12	Card movements
	(Shet et al., 2005)	Indoor	Multi	2 cameras	3	Entry violations, thefts, unattended packages
	(Tran & Davis, 2008)	Outdoor	Multi	640 × 480	n.d.	People entering cars
	(Ijsselmuiden & Stiefelhagen, 2010)	Indoor	Multi	n.d.	5	Individual work, table meeting, presentation, coordinated interaction, standing meeting
	(Khattak et al., 2010)	Indoor	n.d.	n.d.	12	n.d.
	(Pei et al., 2011)	Indoor Outdoor	Multi	n.d.	12	n.d.
	(Song et al., 2013)	Indoor	Single	n.d.	3	Making tea, cocoa, oatmeal
	(Ramirez-Amaro et al., 2013)	Indoor	Single	3 cameras	2	Making a pancake, making a sandwich
D.-b.	(Gupta et al., 2009)	Outdoor	Multi	n.d.	n.d.	Baseball actions
	(Ryoo & Aggarwal, 2009)	Indoor	Multi	320 × 240	Experiment A: 8 Experiment B: 4	Experiment A: two-person interactions; Experiment B: high-level interactions
	(Brendel & Todorovic, 2011)	Outdoor	Multi	Motion	8	Dribbling, jumping, shooting, passing, catching, bouncing, ball trajectory, near rim

3.1. Statistical approaches

Statistical approaches use statistical state-based models to recognize activities allowing an easy representation of probabilities and independencies, providing powerful reasoning mechanisms as well (Borgelt, Gebhardt, & Kruse, 2002). They all share the capability of implementing the structure of a problem without the burden of mathematical details, allowing key dependencies within a problem to be expressed and irrelevancies to be ignored. With enough training data, statistical approaches are able to reliably recognize corresponding activities even in the case of noisy inputs. The need of large training data represents a drawback of these approaches: their combination with syntactic approaches for generating synthetic training data is a viable alternative, which has been explored in object recognition for mobile robots (Pangercic, Tenorth, Jain, & Beetz, 2010; Ruiz-Sarmiento, Galindo, & Gonzalez-Jimenez, 2015).

We found that Bayesian Belief Networks, Probabilistic Petri Nets and Hidden Markov Models are the statistical approaches used in the field of knowledge-based human activity recognition.

Bayesian belief networks. Bayesian belief networks (BBNs) are specifically designed to combine user knowledge and statistical data (Heckerman, 1996; Heckerman, Geiger, & Chickering, 1995). They can be categorized as probabilistic directed acyclic graphical models composed of nodes and arcs. Each node represents a variable of the system. The variables assume different values with a given probability, which might depend on the value of the other nodes. These conditional dependencies are expressed by the arcs of the network. If there is an arc from a node A to a node B of the network, then the probability distribution of the node B depends on the specific value of the node A. The absence of arcs between two nodes represents a mutual independence. The integration of statistical data and knowledge goes as follows. First, the existing knowledge is used to build a BBN. Second, the information extracted from the data update the knowledge, creating a new BBN.

BBNs have been used for modelling activities including context information and prior knowledge in Bauckhage, Hanheide, Wrede, and Sagerer (2004); Gupta and Davis (2007); Hongeng, Nevatia, and Bremond (2004) and Laxton, Lim, and Kriegman (2007).

Usually, the authors take into account also the temporal dependencies between variables, generalizing a BBN into a dynamic BBN.

In Bauckhage et al. (2004) the authors recognize office activity at a desk by using videos and other input sources. They integrate object detection, gesture recognition (tracking of hands) and situation context by means of a BBN that models relations between different hypotheses. In order to guarantee an efficient computation, the lower level modules are distributed over different machines.

In Hongeng et al. (2004) the authors model the events in a video from shape and trajectory features using a hierarchical activity representation. The events are organized into multiple layers of subevents, providing flexibility and modularity in modelling the hierarchy. The context information exploited concerns spatial map and a priori knowledge for subevent expectation. Events performed by multiple actors are recognized by propagating the constraints and the likelihood of event threads in the BBN.

In Gupta and Davis (2007) the authors combine human activity recognition and object detection. Indeed, object context and object reactions can be used to recognize activities, which might otherwise be too similar to distinguish or too difficult to observe. The BBN is used for modelling the object/activity interactions.

A variety of concepts such as object orientation, hierarchy of objects, event/subevents and contextual disambiguation are used in Laxton et al. (2007). They define a multi-level hierarchy of an event and they use a dynamic BBN that models dependencies between states and a temporal probability model over the states.

Probabilistic petri nets. Petri nets are a graphical and mathematical modelling tool to describe and study information processing systems (Rabiner, 1989). They are particularly useful since they permit to express concurrency and to use smart control strategies. Formally, a Petri net is a directed bipartite graph composed of places, transitions, and arcs. In addition, tokens are used in the net to simulate the dynamic activities of the system. Arcs run from a place to a transition or vice versa, and they never directly connect places or transitions. One or more tokens can be situated in each place to determine the state of the system. When a transition is active, a token (if any) moves from its input place to its output place. Some places of the net can be used as terminal places, i.e. when a token reach them, it determines the end of the execution. When the transition activation is determined by a probability distribution, the net becomes a Probabilistic Petri Net (PPN).

In the KBAR framework, PPNs are used in Ghanem, De Menthon, Doermann, and Davis (2004) and Albanese et al. (2008) for representing events composed by subevents that are recognizable by image understanding algorithms.

In Ghanem et al. (2004) the authors proposed a variant of PPN called high-level Petri net. It is an interactive system for querying surveillance video about events where the nets are used as both recognition and representation methods. The queries may not be known in advance and have to be composed from primitive events and previously defined queries. In the recognition part, the input video is preprocessed by applying background subtraction and tracking modules to extract object tracks over time. Object properties and tracks are analysed to detect primitive events that are parts of the final query. The detected primitive events represent inputs to Petri net-based recognition modules, whose function is to recognize a scenario, that is the composition of states and simpler events connected by spatial, temporal or logical relations. In the representation part, the proposed approach describes each event instant by a token. Hence, all instances of the same event are represented by one Petri net and events are represented by tokens in the corresponding net. The advantage of this approach, which extends what introduced in Castel, Chaudron, and Tessier (1996), is that the total number of existing nets is small and fixed.

In Albanese et al. (2008) the authors build one PPN for describing each event. Even if an event may not unfold in the same sequence of the net that represents it, the proposed PPN copes with uncertainty by associating a probability to a particular unfolding of the net. This is achieved by associating a probability to both tokens and transitions. In the beginning, the method assigns a probability score of 1 to every token. Probabilities are then accumulated by multiplying the probabilities of both the tokens and the transitions. When a token reaches the terminal place of the net, the accumulated probability represents the reliability that the particular event occurred in the video.

Hidden Markov models. Hidden Markov Models are state-space models composed of a sequence of states, and they have the property of being able to encode knowledge and model stochastic processes and sequences (Chakrabarti, Rammohan, & Luger, 2007). Similarly to BBNs and PPNs, the knowledge of a domain expert is encoded in the model during the network design.

HMMs were used for KBAR in Duong, Bui, Phung, and Venkatesh (2005); Oliver, Garg, and Horvitz (2004); Robertson and Reid (2006) and Chung and Liu (2008).

Oliver et al. (2004) present a layered HMM. A bottom layer HMM recognizes subevents composing an event. An upper layer HMM treats the recognized subevents as an observation input for modelling the event. This representation is generalized adding multiple layers with different time granularities.

Duong et al. (2005) propose a two-layered extension of the HMM for modelling the inherent hierarchical organization of the activities and their typical duration. The bottom layer represents atomic activities and their duration; the top layer represents a sequence of high-level activities where each high-level activity is made of a sequence of atomic activities. Both multinomial distribution and a discrete Coxian distribution are evaluated for modelling duration.

In Robertson and Reid (2006), the subevents are represented by a feature vector composed of features related to trajectory (position and velocity), and of a set of local motion descriptors. HMMs encoding scene rules are used to smooth a sequence of subevents. Event recognition is achieved by computing the likelihood that an HMM in a set of predefined HMMs models the current subevent sequence. Thus, subevents and events are represented using a hierarchy of abstractions: from subevents with spatio-temporal context to subevent sequences and, finally, the overall event.

In Chung and Liu (2008) the authors define a hierarchical context HMM for behaviour understanding from video streams in a nursing centre. The proposed approach infers elderly behaviours through three modules, which handle spatial context, activity recognition and temporal context, respectively. The activity recognition module leverages the information of spatial context module. The final output is then adjusted by the temporal context module that influences the transition between consecutive states.

3.2. Syntactic approaches

Syntactic approaches define a set of domain-dependent predicates and functions that provide the basis for the statement of facts about the knowledge-base's domain (Brachman & Levesque, 2004). They are characterized by the presence of entities, which are described by types, attributes, functions and relationships. With the vocabulary aforementioned defined, facts of the knowledge base can be expressed by predicates and sentences. Logical reasoning can be used to discover facts that are only implicit in a given knowledge base.

We describe here the syntactic approaches for knowledge exploitation found in the KBAR literature. Although all approaches use, at least implicitly, ontologies and some kind of rules to specify the knowledge to be exploited to perform activity recognition, proposals typically characterize themselves by focusing on a specific technique. In particular, we distinguish methods mainly focused on the use of logic rules, methods that introduce uncertainty and exploit approximate reasoning, methods that use grammars, and methods that specifically address the problem of defining ontologies oriented to human action recognition.

Logic rules. First-order logic is a knowledge base composed of a set of sentences or rules. Rules are constructed using four types of symbols: constants, variables, functions, and predicates. Constants represent objects in the domain of interest. Variables can take the role of any objects in the domain. Functions map a set of objects into a single object. Predicates represent relations among objects in the domain or attributes of objects. First-order logic enables to compactly represent a wide variety of knowledge and permits also automated inference (Robinson, 1965).

Logic rules have been exploited by many KBAR approaches in the literature (Ijsselmuiden & Stiefelhagen, 2010; Ramirez-Amaro et al., 2013; Rincón, Santofimia, & Nebel, 2013; Shet, Harwood, & Davis, 2005).

Shet et al. (2005) propose a system that relies on logic programming to represent composite events performing by multiple actors. Statistical approaches are used to detect primitive events. Then, a high-level reasoning engine recognizes events, which are represented by logical rules between primitives. The approach was validated on a multi-camera surveillance scenario that includes both security and safety violations.

In Ijsselmuiden and Stiefelhagen (2010) the authors use logic rules and the temporal interval relations defined in Allen and Ferguson (1994) to integrate four different information sources, that are, tracking and identification (face recognition and particle filter), visual focus of attention (head pose estimation), gestures (3d body reconstruction by 4 cameras) and speech (who is speaking and what he/she is saying).

In Rincón et al. (2013) the authors propose a two stage methodology which encodes common-sense reasoning in form of logic rules. First, a statistical approach gives an estimate of subevent classification from the video. Second, those results are elaborated to a common-sense reasoning system, which analyses, selects and corrects the initial estimation yielded by the machine learning algorithm. This second stage exploits the three sources of knowledge described in their implementation, i.e. general knowledge, domain specific knowledge and expectation.

Ramirez-Amaro et al. (2013) use first order logic rule to increase the recognition performance of some subevents. The classification problem is split into low level recognition and reasoning. Indeed, they first recognizing three high-level motions (move, not move and tool use) and, second, they use the classification output as input into the reasoning engine to enhance classification of “difficult” low-level activities, such as: reaching, taking, releasing, cutting, sprinkle, etc.

Approximate reasoning. In many practical cases, knowledge is characterized by a certain degree of uncertainty or fuzziness, e.g. the imprecision in the knowledge expressed by human domain experts. In the KBAR literature, we found that both Markov logic networks and fuzzy reasoning have been used. On the one side, Markov logic networks permit to combine probability and first-order logic in a single representation providing the ability to handle uncertainty and tolerate imperfect and contradictory knowledge (Richardson & Domingos, 2006). On the other side, fuzzy reasoning is a process of approximate solution of a set of logic equations managing the non-uniqueness of fuzzy premises (Zadeh, 1975). The result is a formal method to model the human aptitude to manage vague properties (e.g. “small”, “plausible”, “believable”). A simple example is: x is small, x and y are approximately equal, y is more or less small.

Approximate reasoning applications to KBAR can be found in Acampora, Foggia, Saggese, and Vento (2012); Rodriguez-Benitez et al. (2011); Song et al. (2013); Tran and Davis (2008) and Chen et al. (2014).

In Tran and Davis (2008) the authors consider the problem of event modelling and recognition in visual surveillance by introducing an approach based on Markov logic networks. This approach naturally integrates logical reasoning with uncertain analyses produced by computer vision algorithms for object detection, tracking and movement recognition. The knowledge is represented as first-order logic rules, and a heuristic weight is associated to each rule to indicate its confidence measure. These rules are then used in combination with a relaxed deduction algorithm to construct the network.

The approach presented in Rodriguez-Benitez et al. (2011) integrates a vision system, which consists in person segmentation and tracking, with an approximate reasoning algorithm. The authors extract position, velocity and trajectory features as low-level descriptors for the vision system. Then, the data are fuzzified and used as an input for a finite state machine, where a comparison based on a membership matrix for each state is performed to associate a test video to an event.

In Acampora et al. (2012) the authors use temporal features related to human trajectories for classifying a set of predefined subevents. The reliability achieved for each subevent is used as an input vector for a fuzzy system which models an event through an approximate reasoning system composed of a single fuzzy set. Furthermore, additional fuzzy variables are considered in the model for modelling contextual features.

Song et al. (2013) present a general framework for complex event recognition that is well-suited for integrating information that varies in detail and granularity. The system takes as an input objects and subevents as well as descriptions of events extracted from recognized and parsed speech. The system outputs the reconstruction of the event using Markov logic to create a model in which observations can be partial, noisy, and refer to future or temporally ambiguous subevents.

Chen et al. (2014) present a hierarchical recognition framework where knowledge-based logic representation and reasoning are combined with conventional computer vision methods for low-level events classification. Knowledge is coded as rules in the form of logical formulas, and reasoning taking into account uncertainty is used to recognize events with a given credibility degree.

Grammars. Grammars and stochastic grammars present a theoretical basis for modelling structured processes. Context-free grammars (CFGs) are based on a list of primitive variables representing the objects of the model and a set of rules that describes the relations in the model. In a context-sensitive grammar (CSG) the left-hand sides and right-hand sides of any production rules may be surrounded by a context of terminal and nonterminal symbols. Once the rules of a grammar have been formulated, there exists efficient algorithms to parse them (Earley, 1970), which have made them popular in many applications.

In the KBAR framework, grammars were used by Ivanov and Bobick (2000); Moore and Essa (2002) and Pei, Jia, and Zhu (2011).

In Ivanov and Bobick (2000), the lower level detection consists of background subtraction and object tracking. This permits to extract motion features from the identified trajectories and then map the video into a set of discrete subevents. The outputs of this low-level module are the input streams for a stochastic context-free grammar parsing mechanism which provides long range temporal constraints, disambiguates uncertain low-level and includes a priori knowledge.

Moore and Essa (2002) combine visual information with domain-specific information. Each activity event is represented with a unique symbol, allowing for a sequence of interactions to be described as an ordered symbolic string. Then, a model of CFGs, which is developed using underlying rules of an activity, is used to provide the structure for recognizing semantically meaningful behaviour over extended periods. They experimented the recognition of high-level activities on multi-player games, identifying player strategies and behaviour.

Pei et al. (2011) present an event parsing algorithm based on Stochastic CSG for understanding events, inferring the goal of agents, and predicting their plausible intended actions. The CSG represents the hierarchical compositions of events and the temporal relations between the sub-events. The alphabets of the CSG are atomic actions defined by the poses of agents and by their interactions with objects in the scene. The temporal relations are used to distinguish events with similar structures, interpolate missing portions of events, and are learned from the training data.

Ontologies. An ontology can be defined as a formal description of the knowledge within a domain, by means of a set of concepts. Gruber (1993) stated that knowledge in ontologies can be formalized using five components: concepts, relations, functions, axioms and instances, where concepts are usually organized in taxonomies. Ontology popularity is due to their property of capturing domain knowledge in a generic way, thus providing a commonly agreed understanding, which may be shared across applications (Chandrasekaran, Josephson, & Benjamins, 1999).

Ontology-based human activity recognition is an approach for describing the sequence of subevents that is semantically identified with an event. Also entities (actors, objects), environment and interactions between them can be easily included in the ontology formalism. Relevant ontology-based activity recognition systems are presented in Akdemir, Turaga, and Chellappa (2008); Nevatia, Zhao, and Hongeng (2003) and Khattak et al. (2010).

Nevatia et al. (2003) describe human activities with three levels of hierarchy: subsubevents, subevents and event. Subevents are a number of subsubevents with temporal sequencing. Events are a number of subevents with temporal, spatial and logical relationships. This hierarchical event representation naturally leads to a language description of the events that they named as VERL. A heuristic algorithm is designed for the constraint satisfaction problem, recognizing interactions between multiple persons.

In Akdemir et al. (2008) the authors use ontologies to develop a centralized representation of activity that is algorithm-independent. They specify how an event can be constructed using subevent composition and by identifying the role played by each entity in the sequence of subevents. Since events are characterized by complex spatio-temporal interactions between subevents and entities, an event is not considered only as a collection of subevents, but spatio-temporal constraints are included as well.

In Khattak et al. (2010) the activity is recognized through the integration of data coming from sensors and videos. Such data are processed by low-level modules. In case of sensor data it consists of semi-Markov conditional random field that models the activity, its duration and long-transitions between them. In case of videos, low-level processing module provides human body segmentation by using active contour method and bag-of-words classification. The outputs of these two modules are given to ontologies extracting the higher level activity of a set of activities in a series. It is worth noting that in this work logic rules provide intelligent services to the users by analyzing activities managed in domain ontologies, therefore representing an example of how ontologies and logic rules are used within merged frameworks.

3.3. Description-based approaches

Description-based approaches explicitly maintain the spatio-temporal structures of human activities, representing a high-level human activity in terms of temporal, spatial, and logical relationships between the subevents. These approaches cannot be naturally categorized as statistical or syntactic and, further, they allow to describe the complex temporal structures of activities composed of both sequential and concurrent subevents.

Gupta, Srinivasan, Shi, and Davis (2009) proposed a structural model designed to recognize sequential subevents by modelling causality among them. A tree structured AND-OR graph similar to the BBN used in Hongeng et al. (2004) was used to represent composite event of a sports game.

In Ryoo and Aggarwal (2009) the authors present a methodology which describes composite events by integrating statistical recognition techniques from computer vision and knowledge representation concepts from context-free grammar. In the low-level of the system, image sequences are processed to extract poses and gestures. Based on their recognition, the high-level of the system hierarchically recognizes composite actions and interactions occurring in a sequence of image frames. At this level, a CFG semantically represents an event including also spatial and temporal context information, dealing also with uncertainty in low-level event recognition.

Brendel and Todorovic (2011) represent videos by spatio-temporal graphs, where nodes correspond to multi-scale video segments, and edges capture their hierarchical, temporal, and spatial relationships. Given a video, they use a segmentation procedure to obtain homogeneous subsequences, where homogeneity is defined in terms of both pixel intensity and motion properties, at multiple scales. The resulting subsequences are organized in a weighted directed graph, referred to as spatio-temporal graph. From a set of training spatio-temporal graphs of an activity class, the method learns their weighted least squares graph model. Afterwards, in the test phase, a new video is represented by the spatio-temporal graph and classified by matching its graph with the closest activity model in the weighted least squares sense.

In Zhu, Nayak, and Roy-Chowdhury (2013) the authors propose a hierarchical framework that models and recognizes related subevents using motion and context information. The main idea is that the subevents related in space and time rarely occur independently and can serve as the context for each other. Given a video, subevents are automatically detected using motion segmentation based on a nonlinear dynamical model. Afterwards, they merge these segments into activities of interest through a structural model that jointly models the underlying subevents which are related in space and time.

In Li and Fu (2014) the authors present an approach for predicting complex activity by mining temporal sequence patterns. They model the causal relationships between constituent subevent by mining sequential pattern of activities where a series of subevents and context information co-occurrence are encoded as a complex symbolic sequence. They also presented an accumulation function depicting the kind of activities that can be recognized after a few observed subevents against the ones requiring a longer observation.

4. Datasets

In this section we presents public and private datasets used in the KBAR field, whereas a detailed survey on the most relevant public datasets for human action and activity recognition in general can be found in Chaquet, Carmona, and Fernández-Caballero (2013). Table 2 and Table 3 synthetically describe the relevant characteristics of public and private datasets used in the KBAR field, respectively. Besides, we report in the following more details on publicly available datasets to promote their use and to enable the interest readers in finding useful resources.

4.1. Public datasets

Different categorizations exist for public datasets. They can be roughly divided into two groups: in the first one there are the datasets defined for controlled environments, where the KTH dataset (Schuldt et al., 2004) and the Weizmann dataset (Blank, Gorelick, Shechtman, Irani, & Basri, 2005) represent a typical example. They were designed to test general-purpose action recognition systems academically, and they contain videos of different participants performing simple actions such as walking and waving, which are taken by the authors in a controlled environment. Therefore, they are not suited to test KBAR approaches and, for this reason, they are not presented hereinafter. In the second group there are application-oriented datasets obtained from realistic environments (e.g., bank) or from real video media (e.g., TV broadcasts and movies). These repositories contain scenes collected from varying viewpoints with noise and complex backgrounds. However, again not all the datasets in this category have actions more complex than those in the KTH and the Weizmann datasets, and they cannot be used to test KBAR approaches. Similarly to Aggarwal and Cai (1999), we divide the repositories used to measure the performances of KBAR approaches in surveillance datasets and movie datasets.

Surveillance datasets. Surveillance datasets are composed of videos collected in uncontrolled environments, such as airports, banks, roads. The camera viewpoints are similar to those of typical closed-circuit televisions. In the majority od the cases the camera is fixed, so that the backgrounds are static and the scales for persons are mostly constant. Multiple persons and objects appear in the scene simultaneously, and occlusion among them occurs frequently.

The *Bank* dataset (Vu, Bremond, & Thonnat, 2003) contains videos recorded taken in a bank branch, allowing to verify if an algorithm can correctly distinguish between thefts and normal activities.

The *Transport Security Administration* (TSA) dataset (Vaswani, Roy-Chowdhury, & Chellappa, 2005) consists in a video sequence with several airport activities, such as arrival and departure of aircrafts, embarkation and disembarkation of passengers (whose number in the scene varies with time). Different work defines different class activities form this dataset: for instance, in Vaswani et al. (2005) the authors considered passengers deplaning and walking toward the airport terminal as an example of a stationary shape activity, whilst in Akdemir et al. (2008) the authors defined five classes, namely passenger embarkation and disembarkation, aircraft arrival and departure, and luggage cart activity.

The *VIRAT* dataset is a large-scale video repository for assessing the performance of visual event recognition algorithms with a focus on events in outdoor areas with wide coverage (Oh et al., 2011). While some movie datasets consist of short clips showing one action by one individual (Laptev & Pérez, 2007; Liu, Luo, & Shah, 2009), this dataset of surveillance videos consists of many outdoor scenes with actions occurring naturally by non-actors in continuously captured videos of the real world. The dataset is divided into two releases: the first contains 16 event types distributed throughout 25 hours of video, whereas the second is made up of 1 scene 4 hours long. This data is also accompanied by detailed annotations including both moving object tracks and event examples.

PETS dataset is the repository provided at the IEEE International Workshops on Performance Evaluation of Tracking and Surveillance in 2004, 2006, 2007 and 2009 (Ferryman, Crowley, & Shahrokni, 2009; Fisher, 2004). In general, PETS datasets address the problem of group activities such as crowd image analysis within a public space. The PETs 2009 provide researchers to evaluate new or existing detection techniques on dataset captured in a real-world environment. Except for the PETS 2004 repository, the other scenarios were filmed from multiple cameras and involve multiple actors. Table 2 reports information for PETS 2004 dataset (Fisher, 2004) as it has been used in KBAR framework (Acampora et al., 2012; Chen et al., 2014).

The *Inria Xmas Motion Acquisition Sequences* (IXMAS) dataset contains 11 actions, each performed 3 times by 10 actors (5 males / 5 females). Each action is performed by just one actor. In order to provide view-invariance data, the actors freely change their orientation for each acquisition and no further indications on how to perform the actions beside the labels were given (Weinland, Ronfard, & Boyer, 2006).

The *Waiting Room* (WaRo11) dataset (Santofimia, Martinez-del Rincon, & Nebel, 2012) aims at describing the complexity of a real life application with a significant number of complex activities performed by a single actor. To this scope, a specific setup was configured to simulate a waiting room where actions happened without giving any instructions to the subjects. This is facilitated thanks to the presence of several elements interrelated to each other, which may introduce causality and sequentiality as it is found in a real situation. For instance, the presence of a book and a chair could motivate a subject to first pick up the book and then sit down to carry out the action reading.

Movie datasets. Movie datasets are composed of videos obtained from real movie scenes, from TV broadcasts or from YouTube and, therefore, they are not taken in a controlled environment. They differ from the surveillance datasets since camera viewpoints move frequently, and background information is seldom provided.

The *Olympic Sports* dataset (Niebles, Chen, & Fei-Fei, 2010) collects sport activities from YouTube sequences. It contains 16 sport classes, with 50 sequences per class. The sport activities contain complex motions that go beyond simple punctual or repetitive actions. For instance, sequences from the long-jump action class, show an athlete first standing still, in preparation for his/her jump, followed by running, jumping, landing and finally standing up. This contrast with other sport datasets, such as Rodriguez, Ahmed, and Shah (2008) and Ke, Sukthankar, and Hebert (2007), which contains periodic or simple actions like walking, running, golf-swing, ball-kicking and where the major challenge is handling viewpoint and situation-dependent variations rather than recognizing complex structured activities.

The *VISOR* dataset (Vezzani & Cucchiara, 2007) contains a large set of multimedia data and the corresponding annotations. This repository has been conceived as a support tool for different research projects. These data contains physical objects (e.g., people, fixed and mobile objects like trees, vehicles, buildings), action/events (e.g., actions by or interaction between people or events), and context information (metadata, camera information, calibration data, location, date and time). In Table 2 we do not report information about the number of classes and about class description since the large quantity of available data allows researchers to define their experiments, so that it is not possible to depict a experimental unique framework. For instance, in KBAR research Rodriguez-Benitez et al. (2011) defined four scenarios for experimentation. The first one is a vehicle crossing inside a city and in the second one there is a major road communicating two cities, where the authors defined seven prototype actions. The third experiment tries to determine behaviour of pedestrians in a zebra crossing, while the fourth is a traffic scenario.

Table 4
Comparison of knowledge representation and reasoning approaches.

Type	Pros	Cons	Handling uncertainty of low-level modules	Suitable for any low-level video description	Multiple actors
BBN	Combine knowledge and data in a practical way	Large amounts of training data	(Bauckhage et al., 2004; Gupta & Davis, 2007) (Hongeng et al., 2004; Laxton et al., 2007)	(Laxton et al., 2007)	(Hongeng et al., 2004)
PPN	Model concurrency and multiple instances of activities	Automatic structure learning	(Robertson & Reid, 2006) (Albanese et al., 2008)	(Robertson & Reid, 2006) (Albanese et al., 2008)	(Ghanem et al., 2004; Robertson & Reid, 2006) (Albanese et al., 2008)
HMM	Estabilished algorithm for learning and inference	Markovian nature	(Chung & Liu, 2008; Duong et al., 2005)	-	(Chung & Liu, 2008)
Logic rules	Fast and easy inference	Relationships are hard to express	(Rincón et al., 2013; Shet et al., 2005) (Ramirez-Amaro et al., 2013)	(Rincón et al., 2013; Shet et al., 2005)	(Ijsselmuiden & Stiefelhagen, 2010) (Shet et al., 2005)
Ontologies	Standardize activity definitions	Learning ability	(Akdemir et al., 2008; Nevatia et al., 2003)	(Akdemir et al., 2008; Nevatia et al., 2003)	(Akdemir et al., 2008; Khattak et al., 2010) (Nevatia et al., 2003)
Approximate reasoning	Uncertainty handling	Weak membership functions	(Acampora et al., 2012; Rodriguez-Benitez et al., 2011) (Chen et al., 2014; Song et al., 2013; Tran & Davis, 2008)	(Acampora et al., 2012)	(Acampora et al., 2012; Rodriguez-Benitez et al., 2011) (Song et al., 2013; Tran & Davis, 2008)
Grammars	Simplify hierarchical activity definition	Ad hoc definitions	(Ivanov & Bobick, 2000; Moore & Essa, 2002) (Pei et al., 2011)	(Ivanov & Bobick, 2000; Pei et al., 2011)	(Ivanov & Bobick, 2000; Moore & Essa, 2002) (Pei et al., 2011)
Description-based	Maintain the spatio-temporal structures	Ad hoc definitions	(Ryoo & Aggarwal, 2009) (Brendel & Todorovic, 2011)	(Brendel & Todorovic, 2011; Ryoo & Aggarwal, 2009) (Brendel & Todorovic, 2011; Zhu et al., 2013)	(Gupta et al., 2009; Ryoo & Aggarwal, 2009) (Zhu et al., 2013)

5. Discussion

We discuss now the main aspects concerning the emerging KBAR field with the goal of orienting researchers and practitioners when they have to evaluate which is the most suited approach for the problem at hand. Table 4, together with the detailed descriptions presented in previous sections, supports the discussion: it reports the key properties of KBAR approaches and it lists their advantages and disadvantages. The advantages can be summarized as follows. First, the formal frameworks characterizing all methods facilitate an effective exploitation of knowledge and represent therefore an advantage shared by all of them. Indeed, both the ability to specify structural properties of statistical approaches and to explicitly represent rules of syntactic approaches offer to researchers the opportunity of:

- defining in a unified framework any kind of elements related to the recognition of human activities;
- specifying relations and dependencies among these elements, thus allowing the modelling of high-level activities in a flexible way;
- integrating sources of information of different nature and managing easily the uncertainty possibly associated with this knowledge.

Furthermore, we observe specific advantages for each category. In fact, on the one side statistical approaches provide a straightforward way to represent a variety of problems, allowing the expression of the complex conditional dependencies among multiple random variables and deletion of irrelevant links between them. On the other side, syntactic approaches provide excellent tools for modelling and exploiting the knowledge of a particular domain such as the set of domain-dependent entity descriptions, predicates and functions. Description-based methods try somehow to get the best of both categories modelling jointly the event structure and context knowledge.

Turning our attention to the limitations of the described approaches, we point out what follows. In case of statistical approaches, both dynamic BBNs and HMMs make the assumption of Markovian dynamics, i.e. the conditional probability distribution of future subevents depends only upon the present subevent. This restricts their applicability to relatively simple and stationary temporal sequences. Moreover, whilst the problem of learning the structure of a PPN from training data has not yet been formally addressed, BBNs and HMMs require large amount of training data for learning conditional probability densities. It is worth noting that a possible solution to this issue has been explored in a different but related task, i.e., object recognition for mobile robots: in this case, there exist efforts combining statistical approaches with ontologies of concepts for generating synthetic training data (Pangercic et al., 2010; Ruiz-Sarmiento et al., 2015). With respect to syntactic approaches, a main drawback has regarded the fact that the symbolic definition of activities has to be constructed in an empirical manner. For instance, both rules of a grammar and sets of logic rules are usually manually defined. This tends to limit the applicability of these rules, and makes them well suited only for the specific applications for which they have been designed. Furthermore, syntactic approaches are unsuited for activities composed of concurrent subevents: this happens since the temporal ordering adopted to model a high-level activity as a string of subevents has to be strictly sequential.

A researcher can be interested in comparing the performances of these methods: however, as is apparent from tables presented in section 4, the use of benchmark datasets is not diffused. Indeed, most of the datasets used to measure the recognition accuracy are private, each with different peculiarities, and also public datasets have been tested only in a few cases on more than one method. Moreover, in some cases system performance is not even clearly reported (Akdemir et al., 2008; Albanese et al., 2008; Ghanem et al., 2004; Hongeng et al., 2004; Ivanov & Bobick, 2000; Khattak et al., 2010; Nevatia et al., 2003;

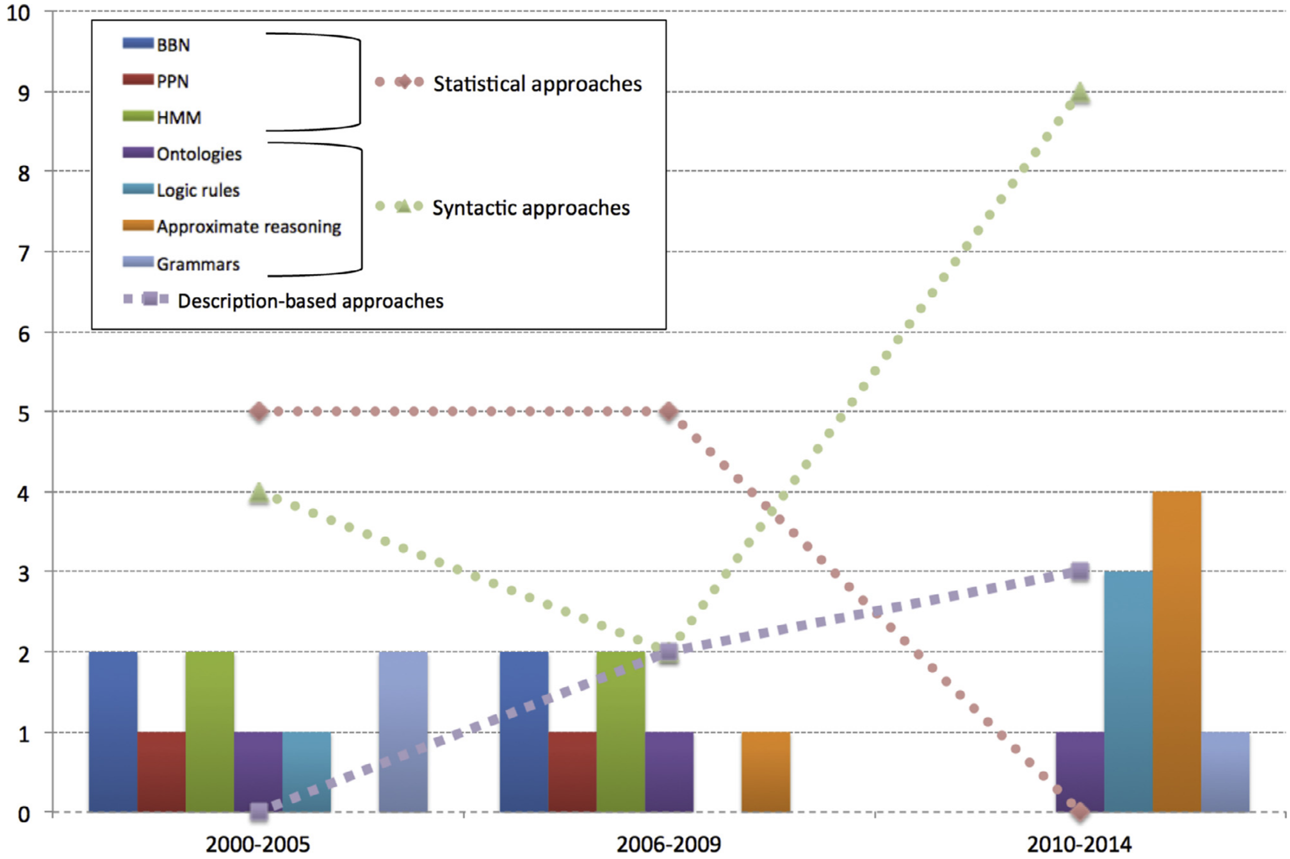

Fig. 3. Number of papers published over the years for each category.

Ramasso, Rombaut, & Pellerin, 2006; Robertson & Reid, 2006; Rodriguez-Benitez et al., 2011; Shet et al., 2005; Tran & Davis, 2008; Zhu et al., 2013). This makes a comparison among methods performance infeasible.

Nevertheless, we found a quantitative element emerging from the reported work: while most of the methods propose a KBAR system in a scenario where a single-layered approach could not even be usable, some papers explicitly show that recognition rate increases when contextual features and a priori knowledge is exploited (Gupta & Davis, 2007; Li & Fu, 2014; Rincón et al., 2013; Rodriguez-Benitez et al., 2011; Song et al., 2013).

Further to the comparison presented in Table 4, Fig. 3 shows the total number of papers belonging to a given approach that were published in the last 15 years, grouped in three temporal intervals. We observe that syntactic approaches are gradually establishing themselves as the most popular for knowledge-based human activity analysis. We speculate that reasons behind this trend could be mostly related to the ability of rules to describe the large variety of issues involved in the recognition of human activities. Moreover, we deem that a deciding factor in establishing this trend is the recent development of designing tools for syntactic approaches (e.g. Nevatia, Hobbs, & Bolles, 2004), which makes now easier to produce a task-independent knowledge-base, while maintaining at the same time the opportunity to make task-dependent simplifications. As long as standardization of activity definitions through a general representation framework become feasible, knowledge described with syntactic methods turns out to be portable to different scenarios, so enabling interoperability of different approaches and making easier the comparison of system performance (Akdemir et al., 2008). A different consideration that may explain the popularity of rule-based approaches is the growing interest for applications, i.e. ADL recognition and video-surveillance, where the activities of interest are limited in number and strongly standardized. Indeed, the way a person sets the table or a subject attempts to sneak in a restricted area can be easily defined and represented by the aforementioned designing tools for semantic and syntactic approaches.

6. Future directions and conclusions

There is a number of methods available for human activity recognition in video streams by means of high level representation, contextual information and reasoning methods, often combined with low-level statistical approaches. Concretely, the 30 contributions surveyed here have been categorized into statistical, syntactic and description-based approaches.

Although significant improvements have been recently achieved by KBAR, this review has shown that further research efforts are still needed. In this respect, we propose the following main directions, which may improve activity recognition performance and the ability to better evaluate these improvements.

Improving low-level recognition. Significant efforts have been made regarding the extraction of low-level features for the recognition of simple activities (Poppe, 2010). Nevertheless, global features like colour and texture as well as the more recent local approaches based on interest points detection and bag-of-words classification are still negatively affected by factors characterizing complex activities, such as noise, occlusions, interactions, etc. Weaknesses in the extraction of robust features from the raw video stream can easily propagate to higher levels and impair the knowledge-based reasoning. In addition to the visual features, KBAR approaches could easily include different sources of data (e.g. sensors of other nature): while they may help the high level reasoning, they too may be negatively affected by real-world conditions, requiring to look for effective trade-offs.

Scene understanding. In this review we reported several examples of approaches including context information supporting the recognition of high-level activities, typically not immediately related to low-level human motion. However, a complete scene understanding and information on human-environment interactions are still poorly exploited. Indeed, besides recognizing and tracking interacting objects, also non-interacting objects can help scene understanding and provide a useful support for the recognition of high-level activities. Similarly knowledge of the actor state (head pose, facial expression, distance from specific objects, etc.) can be of great help.

Deploying the knowledge base. A clear need of research in activity recognition is the development of a comprehensive and easily accessible labelled data on human activity videos with different characteristics in terms of application contexts, acquisition modalities, levels of complexity, etc. In a KBAR perspective, this data should include non-visual information describing the a priori knowledge on activities of interest. Moreover, although KBAR statistical methods need for large training sets to build the probabilistic model, it is still a challenge to make available a knowledge base that effectively replaces, or even substantially integrates, information extracted from the visual data. Exploring how to combine statistical and syntactic approaches to generate synthetic data to overcome the need of large training data is a topic that deserve more efforts, as it has happened in other related fields (e.g., object recognition in mobile robots).

Benchmarks and performance assessment. Several public datasets have been introduced in the last fifteen years, encouraging researchers to explore various recognition directions in human action and activity recognition. The use of publicly available datasets has three main advantages. First, they save time and resources, that is, there is no need to record new video-sequences or pay for them, so researchers can direct their efforts towards the algorithms and the implementations. Second, their use focuses the research target permitting great advancements in the field. Third, and this is even the more important advantage, the use of the same datasets facilitates the comparison of different approaches and gives insight into the abilities of the different methods. Unfortunately, in the field of KBAR systems we observed that the use of public dataset is very reduced. We deem that this happens as there is not any universally recognized benchmark. Indeed, even if there are publicly available datasets, e.g. the surveillance datasets provided at the PETS workshops on 2004, 2006, 2007, they could be used at least to evaluate performance in a specific application context. Furthermore, we also noticed that there is not a uniform protocol (performance metrics, training data) that would permit to quantitatively compare different approaches. The use of benchmark repositories, the definition of common experimental frameworks and the adoption of uniform performance measures therefore represent a challenge in KBAR research.

Conflict of interests

None declared.

Acknowledgements

The authors thank the anonymous reviewers that provided many useful suggestions to improve the first version of the manuscript.

References

Acampora, G., Foggia, P., Saggese, A., & Vento, M. (2012). Combining neural networks and fuzzy systems for human behavior understanding. In *Advanced video and signal-based surveillance (AVSS), 2012 IEEE ninth international conference on* (pp. 88–93). IEEE.

Aggarwal, J., & Ryoo, M. S. (2011). Human activity analysis: a review. *ACM Computing Surveys (CSUR), 43*(3), 16.

Aggarwal, J. K., & Cai, Q. (1999). Human motion analysis: a review. *Computer Vision and Image Understanding, 73*(3), 428–440.

Akdemir, U., Turaga, P., & Chellappa, R. (2008). An ontology based approach for activity recognition from video. In *Proceedings of the 16th ACM international conference on multimedia* (pp. 709–712). ACM.

Albanese, M., Chellappa, R., Moscato, V., Picariello, A., Subrahmanian, V., Turaga, P., et al. (2008). A constrained probabilistic petri net framework for human activity detection in video. *Multimedia, IEEE Transactions on, 10*(6), 982–996.

Allen, J. F., & Ferguson, G. (1994). Actions and events in interval temporal logic. *Journal of Logic and Computation, 4*(5), 531–579.

Bauckhage, C., Hanheide, M., Wrede, S., & Sagerer, G. (2004). A cognitive vision system for action recognition in office environments. In *Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on: vol. 2* (pp. II–827). IEEE.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., & Basri, R. (2005). Actions as space–time shapes. In *Computer vision, 2005. ICCV 2005. Tenth IEEE international conference on: vol. 2* (pp. 1395–1402). IEEE.

Bobick, A. F. (1997). Movement, activity and action: The role of knowledge in the perception of motion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 352*(1358), 1257–1265.

Borgelt, C., Gebhardt, J., & Kruse, R. (2002). Graphical models. In *Proceedings of international school for the synthesis of expert knowledge (ISSEK'98)*. Citeseer.

Brachman, R. J., & Levesque, H. J. (2004). *Knowledge representation and reasoning*. Morgan Kaufmann.

Brendel, W., & Todorovic, S. (2011). Learning spatiotemporal graphs of human activities. In *Computer vision (ICCV), 2011 IEEE international conference on* (pp. 778–785). IEEE.

Cardinaux, F., Bhowmik, D., Abhayaratne, C., & Hawley, M. S. (2011). Video based technology for ambient assisted living: A review of the literature. *Journal of Ambient Intelligence and Smart Environments, 3*(3), 253–269.

Castel, C., Chaudron, L., & Tessier, C. (1996). What is going on? A high level interpretation of sequences of images. In *ECCV workshop on conceptual descriptions from images* (pp. 13–27).

Chakrabarti, C., Rammohan, R., & Luger, G. F. (2007). Diagnosis using a first-order stochastic language that learns. *Expert Systems with Applications, 32*(3), 832–840.

Chandrasekaran, B., Josephson, J. R., & Benjamins, V. R. (1999). What are ontologies, and why do we need them? *Intelligent Systems and Their Applications, IEEE, 14*(1), 20–26.

Chang, S.-F. (2002). The holy grail of content-based media analysis. *Multimedia, IEEE, 9*(2), 6–10.

Chaquet, J. M., Carmona, E. J., & Fernández-Caballero, A. (2013). A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding, 117*(6), 633–659.

Chen, S., Clawson, K., Jing, M., Liu, J., Wang, H., & Scotney, B. (2014). Uncertainty reasoning based formal framework for big video data understanding. In *Web intelligence and intelligent agent technologies, 2014. IEEE/WIC/ACM joint conference on* (pp. 487–494).

Choi, J., Cho, Y.-i., Han, T., & Yang, H. S. (2008). A view-based real-time human action recognition system as an interface for human computer interaction. In *Virtual systems and multimedia* (pp. 112–120). Springer.

Chung, P.-C., & Liu, C.-D. (2008). A daily behavior enabled hidden markov model for human behavior understanding. *Pattern Recognition, 41*(5), 1572–1580.

Crevier, D., & Lepage, R. (1997). Knowledge-based image understanding systems: a survey. *Computer Vision and Image Understanding, 67*(2), 161–185.

Dey, A. K. (2001). Understanding and using context. *Personal and ubiquitous computing, 5*(1), 4–7.

Dimitrova, N. (2003). Multimedia content analysis: The next wave. In *Image and video retrieval* (pp. 9–18). Springer.

Duong, T. V., Bui, H. H., Phung, D. Q., & Venkatesh, S. (2005). Activity recognition and abnormality detection with the switching hidden semi-markov model. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on: vol. 1* (pp. 838–845). IEEE.

Earley, J. (1970). An efficient context-free parsing algorithm. *Communications of the ACM, 13*(2), 94–102.

Ferryman, J., Crowley, J., & Shahrokni, A. (2009). Performance Evaluation of Tracking and Surveillance, IEEE International Workshop on. http://www.cvg.reading.ac.uk/PETS2009/a.html.

Fisher, R. B. (2004). The PETS04 surveillance ground-truth data sets. In *Performance evaluation of tracking and surveillance, 2004. 6th IEEE International workshop on* (pp. 1–5).

Ghanem, N., De Menthon, D., Doermann, D., & Davis, L. (2004). Representation and recognition of events in surveillance video using petri nets. In *Computer vision and pattern recognition workshop, 2004. CVPRW'04. conference on* (p. 112). IEEE.

Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R. (2007). Actions as space–time shapes. *Transactions on Pattern Analysis and Machine Intelligence, 29*(12), 2247–2253.

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition, 5*(2), 199–220.

Gupta, A., & Davis, L. S. (2007). Objects in action: An approach for combining action understanding and object perception. In *Computer vision and pattern recognition, 2007. cvpr'07. IEEE conference on* (pp. 1–8). IEEE.

Gupta, A., Srinivasan, P., Shi, J., & Davis, L. S. (2009). Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos. In *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on* (pp. 2012–2019). IEEE.

Hanjalic, A., Lienhart, R., Ma, W.-Y., & Smith, J. R. (2008). The holy grail of multimedia information retrieval: So close or yet so far away? *Proceedings of the IEEE, 96*(4), 541–547.

Heckerman, D. (1996). Bayesian networks for knowledge discovery. *Advances in knowledge discovery and data mining, 11*, 273–305.

Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning, 20*(3), 197–243.

Hongeng, S., Nevatia, R., & Bremond, F. (2004). Video-based event recognition: activity representation and probabilistic recognition methods. *Computer Vision and Image Understanding, 96*(2), 129–162.

Ijsselmuiden, J., & Stiefelhagen, R. (2010). Towards high-level human activity recognition through computer vision and temporal logic. In *KI 2010: Advances in artificial intelligence* (pp. 426–435). Springer.

Ivanov, Y. A., & Bobick, A. F. (2000). Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 22*(8), 852–872.

Ke, Y., Sukthankar, R., & Hebert, M. (2007). Spatio-temporal shape and flow correlation for action recognition. In *Computer vision and pattern recognition (CVPR), 2007. IEEE Conference on* (pp. 1–8). IEEE.

Khan, Z. A., & Sohn, W. (2011). Abnormal human activity recognition system based on r-transform and kernel discriminant technique for elderly home care. *Consumer Electronics, IEEE Transactions on, 57*(4), 1843–1850.

Khattak, A. M., The Vinh, L., Hung, D. V., Truc, P. T., Hung, L. X., Guan, D., et al. (2010). Context-aware human activity recognition and decision making. In *e-Health networking applications and services (Healthcom), 2010 12th IEEE international conference on* (pp. 112–118). IEEE.

Kruger, V., Kragic, D., Ude, A., & Geib, C. (2007). The meaning of action: A review on action recognition and mapping. *Advanced Robotics, 21*(13), 1473–1501.

Laptev, I., & Pérez, P. (2007). Retrieving actions in movies. In *Computer vision (ICCV), 2007. IEEE 11th international conference on* (pp. 1–8). IEEE.

Laxton, B., Lim, J., & Kriegman, D. (2007). Leveraging temporal, contextual and ordering constraints for recognizing complex activities in video. In *Computer vision and pattern recognition, 2007. CVPR'07. IEEE conference on* (pp. 1–8). IEEE.

Li, K., & Fu, Y. (2014). Prediction of human activity by discovering temporal sequence patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 36*(8), 1644–1657.

Lin, W., Sun, M.-T., Poovandran, R., & Zhang, Z. (2008). Human activity recognition for video surveillance. In *Circuits and systems (ISCAS), 2008. IEEE international symposium on* (pp. 2737–2740). IEEE.

Liu, J., Luo, J., & Shah, M. (2009). Recognizing realistic actions from videos "in the wild". In *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on* (pp. 1996–2003). IEEE.

McKenna, T. (2003). Video surveillance and human activity recognition for anti-terrorism and force protection. In *Advanced video and signal based surveillance, 2003. IEEE conference on* (p. 2). IEEE.

Moore, D., & Essa, I. (2002). Recognizing multitasked activities from video using stochastic context-free grammar. In *AAAI/IAAI* (pp. 770–776).

Nevatia, R., Hobbs, J., & Bolles, B. (2004). An ontology for video event representation. *Computer vision and pattern recognition workshop, 2004. CVPRW'04. conference on*. IEEE. 119–119

Nevatia, R., Zhao, T., & Hongeng, S. (2003). Hierarchical language-based representation of events in video streams. *Computer vision and pattern recognition workshop, 2003. CVPRW'03. conference on*: vol. 4. IEEE. (pp. 39–39).

Niebles, J. C., Chen, C.-W., & Fei-Fei, L. (2010). Modeling temporal structure of decomposable motion segments for activity classification. In *Computer vision–ECCV 2010* (pp. 392–405). Springer.

Nigro, H. O., Císaro, S. E. G., & Xodo, D. H. (2008). *Data mining with ontologies: Implementations, findings, and frameworks*. Information Science Reference.

Niu, W., Long, J., Han, D., & Wang, Y.-F. (2004). Human activity detection and recognition for video surveillance. In *Multimedia and Expo (ICME), 2004. IEEE International Conference on: 1* (pp. 719–722). IEEE.

Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J. T., et al. (2011). A large-scale benchmark dataset for event recognition in surveillance video. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on* (pp. 3153–3160). IEEE.

Oliver, N., Garg, A., & Horvitz, E. (2004). Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding, 96*(2), 163–180.

Pangercic, D., Tenorth, M., Jain, D., & Beetz, M. (2010). Combining perception and knowledge processing for everyday manipulation. In *Intelligent robots and systems (IROS), 2010. IEEE/RSJ international conference on* (pp. 1065–1071). IEEE.

Pei, M., Jia, Y., & Zhu, S.-C. (2011). Parsing video events with goal inference and intent prediction. In *Computer vision (ICCV), 2011 IEEE international conference on* (pp. 487–494). IEEE.

Pentland, A. (1998). Smart rooms, smart clothes. In *Pattern recognition, 1998. proceedings. fourteenth international conference on: vol. 2* (pp. 949–953). IEEE.

Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing, 28*(6), 976–990.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE, 77*(2), 257–286.

Ramasso, E., Rombaut, M., & Pellerin, D. (2006). A temporal belief filter improving human action recognition in videos. In *Acoustics, speech and signal processing, 2006. ICASSP 2006 proceedings. 2006 IEEE international conference on: vol. 2* (p. II). IEEE.

Ramirez-Amaro, K., Kim, E., Kim, J., Zhang, B., Beetz, M., & Cheng, G. (2013). Enhancing human action recognition through spatio-temporal feature learning and semantic rules. In *RAS, 2013. IEEE Conference on*. IEEE.

Richardson, M., & Domingos, P. (2006). Markov logic networks. *Machine Learning, 62*(1-2), 107–136.

Rincón, J. M. d., Santofimia, M. J., & Nebel, J.-C. (2013). Common-sense reasoning for human action recognition. *Pattern Recognition Letters, 34*(15), 1849–1860.

Robertson, N., & Reid, I. (2006). A general method for human activity recognition in video. *Computer Vision and Image Understanding, 104*(2), 232–248.

Robinson, J. A. (1965). A machine-oriented logic based on the resolution principle. *Journal of the ACM (JACM), 12*(1), 23–41.

Rodriguez, M. D., Ahmed, J., & Shah, M. (2008). Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2008. IEEE Conference on* (pp. 1–8). IEEE.

Rodriguez-Benitez, L., et al. (2011). Approximate reasoning and finite state machines to the detection of actions in video sequences. *International Journal of Approximate Reasoning, 52*(4), 526–540.

Rohrbach, M., Amin, S., Andriluka, M., & Schiele, B. (2012). A database for fine grained activity detection of cooking activities. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 1194–1201). IEEE.

Ruiz-Sarmiento, J.-R., Galindo, C., & Gonzalez-Jimenez, J. (2015). Exploiting semantic knowledge for robot object recognition. *Knowledge-Based Systems, 86*, 131–142.

Ryoo, M. S., & Aggarwal, J. K. (2009). Semantic representation and recognition of continued and recursive human activities. *International Journal of Computer Vision, 82*(1), 1–24.

Santofimia, M. J., Martinez-del Rincon, J., & Nebel, J.-C. (2012). Common-sense knowledge for a computer vision system for human action recognition. In *Ambient assisted living and home care* (pp. 159–166). Springer.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local svm approach. In *Pattern Recognition (ICPR), 2004. International Conference on: vol. 3* (pp. 32–36). IEEE.

Shang, Y., & Lee, E.-J. (2011). Human multimedia display interface based on human activity recognition. In *3rd international conference on digital image processing* (p. 80091Y). International Society for Optics and Photonics.

Shet, V. D., Harwood, D., & Davis, L. S. (2005). Vidmap: video monitoring of activity with prolog. In *Advanced video and signal based surveillance, 2005. AVSS 2005. IEEE conference on* (pp. 224–229). IEEE.

Song, Y. C., Kautz, H., Allen, J., Swift, M., Li, Y., Luo, J., et al. (2013). A markov logic framework for recognizing complex events from multimodal data. In *Proceedings of the 15th ACM on international conference on multimodal interaction* (pp. 141–148). ACM.

Suriani, N. S., Hussain, A., & Zulkifley, M. A. (2013). Sudden Event Recognition: A Survey. *Sensors, 13*, 9966–9998.

Tran, S. D., & Davis, L. S. (2008). Event modeling and recognition using markov logic networks. In *Computer vision–ECCV 2008* (pp. 610–623). Springer.

Turaga, P., Chellappa, R., Subrahmanian, V. S., & Udrea, O. (2008). Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on, 18*(11), 1473–1488.

Vaswani, N., Roy-Chowdhury, A. K., & Chellappa, R. (2005). "shape activity": a continuous-state HMM for moving/deforming shapes with application to abnormal activity detection. *Image Processing, IEEE Transactions on, 14*(10), 1603–1616.

Vezzani, R., & Cucchiara, R. (2007). Visor: Video surveillance online repository. *BMVA symposium on Security and surveillance: performance evaluation*.

Vishwakarma, S., & Agrawal, A. (2013). A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer, 29*(10), 983–1009.

Vu, V.-T., Bremond, F., & Thonnat, M. (2003). Automatic video interpretation: A novel algorithm for temporal scenario recognition. In *IJCAI: 3* (pp. 1295–1300).

Wang, L., Hu, W., & Tan, T. (2003). Recent developments in human motion analysis. *Pattern Recognition, 36*(3), 585–601.

Weinland, D., Ronfard, R., & Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding, 104*(2), 249–257.

Zadeh, L. A. (1975). Fuzzy logic and approximate reasoning. *Synthese, 30*(3-4), 407–428.

Zhu, Y., Nayak, N. M., & Roy-Chowdhury, A. K. (2013). Context-aware activity recognition and anomaly detection in video. *Selected Topics in Signal Processing, IEEE Journal of, 7*(1), 91–101.

Ziaeefard, M., & Bergevin, R. (2015). Semantic human activity recognition: A literature review. *Pattern Recognition, 48*, 2329–2345.

Zouba, N., Boulay, B., Bremond, F., & Thonnat, M. (2008). Monitoring activities of daily living (ADLs) of elderly based on 3D key human postures. In *Cognitive vision* (pp. 37–50). Springer.